IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.15, No.3, July 2021, pp. 255~264
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.67392

Covid-19 Hoax Detection Using KNN in Jaccard Space

Ema Utami*[1], Ahmad Fikri Iskandar[2], Wahyu Hidayat[3], Agung Budi Prasetyo[4], Anggit Dwi Hartanto[5]
[1,2,3,4,5] Magister Teknik Informatika, Univeristas Amikom Yogyakarta, Yogyakarta, Indonesia
e-mail: ***[1]ema.u@amikom.ac.id**, [2]andfikri@gmail.com, [3]wahyuhublaa@gmail.com, [4]agung.bp@excelindo.co.id, [5]anggit@amikom.ac.id

Abstrak

Media sosial telah menjadi kunci komunikasi untuk memicu pemikiran, dialog, dan tindakan seputar masalah sosial. Hoaks adalah suatu informasi yang ditambahi atau dikurangi isi dari berita yang sebenarnya terjadi. Seperti halnya berita penyebaran Covid-19 yang belum dipastikan kebenarannya, sehingga dapat menimbulkan kekhawatiran dimasyarakat. Tujuan dari penelitian ini adalah melakukan modifikasi model KNN dengan Jaccard Space dalam klasifikasi berita hoaks terkait Covid-19. Data yang digunakan berasal dari Jabar Saber Hoaks dan Jala Hoaks. Hasil klasifikasi dengan KNN dengan Jaccard Space dengan stemming Nazief & Adriani mendapatkan akurasi tertinggi dari model lain pada peneltiian ini. Akurasi pada model KNN pada Jaccard Space dengan dengan stemming Nazief & Adriani dan K=5 sebesar 75,89%, sedangkan untuk Naïve Bayes sebesar 65,18%.

Kata kunci*— Hoax, Covid-19, KNN, Jaccard, Nazief & Adriani*

Abstract

Social media has become a communication key to spark thinking, dialogue and action around social issues. Hoax is information that added or subtracted from the content of the actual news. The spread of unconfirmed Covid-19 news can cause public concern. The purpose of this research was to modify KNN with Jaccard Space in the classification of hoax news related to Covid-19. The data used from Jabar Saber Hoaks and Jala Hoaks. The classification results with KNN with Jaccard Space and stemming Nazief & Adriani get the highest accuracy than other models in this research. The accuracy of the KNN model on the Jaccard Space with stemming Nazief & Adriani and K = 5 was 75.89%, while for Naïve Bayes was 65.18%.

Keywords— *Hoax, Covid-19, KNN, Jaccard, Nazief & Adriani*

1. INTRODUCTION

1.1 Background

Hoax is information that added from the content of actual news [1]. The element of manipulation or modification in the news is often used to respond so that the news will go viral. Indonesian hoax news phenomenon is seen as causing various problems. This happens because of the rapid spread of hoax news on social media which is the basis for communication between users without checking the news first. [2]. A news is said to be true if the reader has proven the truth of the news content.

Jabar Saber Hoaks [3] listed levels from lowest to highest in misinformation and disinformation hoax news found on social media namely:

a. *Satire or Parody,*
b. *False Connection,*

Received July 6th,2021; Revised July 27th, 2021; Accepted July 30th, 2021

c. *Misleading Content,*
d. *False Context,*
e. *Imposter Content,*
f. *Manipulated Content,*
g. *Fabricated Content*

Hoax news was found on news portals [4] social media [5] and clickbait news [6]. Machine learning models used by researchers in detecting hoax news, such as Naive Bayes [4], SVM [5], and KNN [6]. KNN uses distance value *d(a,b)* between the training data and testing data with parameter K (number of neighbor) [6] [7]. Euclidean distance performance in KNN has not shown its best performance. One of the most widely used is Jaccard. The distance value is obtained by compating and calcuating the similarity value of two documents. Jaccard coefficient looks up which words are the same divided by the total documents [8].

Nazief & Adriani algorithm [9] affects the results with TF-IDF (Term Frequency-Inverse Document Frequency) process because the word converted into a root word. This implementation that has been carried out in the classification process using the Naïve Bayes can optimize word extraction in the data obtained and affect the results of N-Gram [10]. This stemming has rule of deleting words affixed into standard words.

Based on the explaination above, this research aims to develop a model that can predict level of hoax news using modified KNN in Jaccard space with text preprocessing Stemming Nazief & Adriani. This modified model was applied to find most similar hoax data, and expected to increase KNN model performance.

1.2 Previous Works

Naïve Bayes model with modified TF-IDF have been used to classify indonesian hoax news by Mustofa [11]. The data used is 360 documents. In accordance with the testing with the cross validation (CV) 10 times, it was achieved 85% of accuracy with CV equal 6 and error rate of 15%. Hoax detection from the news applies Naïve Bayes and Cosine similarity [4]. Best performance detection using Naïve Bayes model achieved 91% precision, 100% recall, and 95% f-measures.

Clickbait news classification [6] with three scenarios KNN with euclidean distance and the value of parameter K ranging from 1 to 15 with data of 1000 news published from January 2020.. The scenario used a combination number of training and testing data (80:20, 50:50 and 20:80). Accuracy KNN with K=11 is 71% with number of training and testing data 80:20. Fake news research used data from 2282 Facebook posts has not achieved best performance. [7]. This facebook posts consist of 1669 posts have the label "Mostly True", 264 posts have the label "No Factual", 245 has the label "Mixture of True and False" and 104 posts have the label "Mostly False". Accuarcy of KNN model obtained accuracy of 79%. Hoax news on realated Covid-19 in group chat UNNES [12] with hoax data 500, message 4519 and media contained in the message 1435. Classifier of this scenario is KNN with K=1 obtained average accuracy 54% with minimum accuracy of 14% and a maximum accruacy of 91%.

Research on hoax news detection [13] using the SVM model. Accuracy obtained using TF-IDF is 85%, with 90% for Non-Hoax labels and 80% for hoax labels. The potential for spreading hoaxes on Twitter social media [5] with TF-IDF and SVM models. There are two test scenarios carried out, namely in the first scenario using data sharing variants, namely 90:10, 80:20 and 50:50, while for the second scenario by comparing the existing features to determine the effect of features. Best accuracy obtained by first scenario is 78% accuracy with 90:10 training data and testing data.

Research related to the detection of fake news against the Covid-19 disease with C4.5 Decision Tree Model [14] and combination of N-Gram and TF-IDF. The data used social media twitter hoax against the corona virus, politics and the environment with the hoax label as 49.4% and 50.6% non-hoax. The best accuracy was 72.04% obtained by Unigram with 90:10 training and testing data.

Implementation of the Jaccard Index has been conducted to determine the similarity index in the classification of ham and spam e-mails [15]. The method used is the Document Similarity Index (DSI) which is calculated from the Jaccard Index and gets 98% precision results for Ham labels and 98% for Spam labels. Improving truth detection in social media using Scalable and Robust Truth Discovery (SRTD) count with Jaccard Similarity has done by Sangwan [16]. Keys point focus of this resarch are attitude, uncertainty and independent score can be determined by using WORDNET programmed in Java. This scenario used to enhance the exploration of reality through similar terms.

Table 1 Related Works of Hoax Detection

Method	***Result***	**Limitation**
Hoax News		
Naive Bayes [4]	Scenario URL, best precision, recall and f-measure are 91%, 100%,and 95%.	This research not compare the performance of NB with several scenarios.
Naïve Bayes [11]	Best accuracy obtained for classifying hoax news using NB is 85% with CV 6.	This research not compare the performance of NB with several scenarios.
SVM [13]	Average accuracy of SVM model testing has an accuracy of 85%.	Testing stage is not explained detail calculation in the testing used.
KNN [6]	Classification of clickbait news with KNN obtained an accuracy of 71% at K = 11 with 80:20 data sharing.	Only focuses on 1 model (KNN) with euclidean distance, so that the accuracy performance is not optimal.
Hoax in Social Media		
KNN [7]	Classification fake news results using data from Social Media Facebook with the KNN model obtained an accuracy of 79%.	There is no modification to the KNN model.
SVM [5]	In the first scenario, the results of this research obtained great accuracy with an accuracy rate of 78% (90:10 training data and testing data)	The limitations of this research is using the unbalanced amount of data on each label, namely 67% for Hoax data labels and 33% for Non-Hoax label data.
KNN[12]	KNN model with K=1 obtained average accuracy 54% with minimum accuracy of 14% and a maximum accruacy of 91% on data related Covid-19.	There is no modification to the KNN model.
Decision Tree [14]	Best accuracy of N-Gram combination obtained 72,04% (Unigram) with 90:10 training and testing data related Covid-19.	There are no model comparison.
Jaccard Similarity [16]	This scenario using WORDNET can improve the score that helps to determine truth better.	This reaseach is not compare the similarity with euclidean, manhataan or others.
Mail Spam		
Jaccard Similarity [15]	The result of this resarch is 98% precision for ham and spam labels.	The limitations of the research were not comparing other similarity methods.

According Table 1, this research conduct using scenarios with and without stemming Nazief & Adriani to classify hoax detection using Modified KNN and Naïve Bayes. The main focus of this research compare the results of KNN classification with modified KNN with Jaccard Space and stemming Nazief & Adriani in the classification of hoax news related to Covid-19.

In summary, contributions of this work processed and classify hoax news related to Covid-19 using modified KNN in Jaccard space with Stemming Nazief & Adriani from Jabar Saber Hoaks and Jala Hoaks. This research organized as follows, Section 1 discusses the background, hoax concepts, and recent research about hoax detection. Section 2 explain

classification method used. Section 3 presents classification result. Section 4 concludes this work and future research.

2. METHODS

This section presents the dataset, data preprocessing, modeling KNN with Jaccard Space, and the scenario of evaluation.

2.1 Data

Data collection uses web crawler by extracting and processing in text information on a web page. Thus, information placed in index based on keywords to csv file. Web pages to be executed by crawlers are Jabar Saber Hoaks and Jala Hoaks (sample on Table 2) :

Table 2 Sample data

News	Fact	Conclusion	Category	Class
Beredar pesan berantai di aplikasi WhatsApp yang menginformasikan pengumuman vaksin Covid-19 untuk lansia dan bukan lansia di Puskesmas Kecamatan Kramat Jati dengan hanya membawa e-KTP asli dan memakai masker dengan protokol 3M.	Berdasarkan hasil koordinasi Tim Jalahoaks dengan Dinas Kesehatan Provinsi DKI Jakarta (17/03/2021), diperoleh klarifikasi bahwa Puskesmas Kecamatan Kramat Jati tidak pernah mengeluarkan pengumuman tersebut. Adapun nomor hotline Puskesmas Kecamatan Kramat Jati dapat dihubungi melalui nomor 0895321748470.	Informasi tentang pengumuman vaksin Covid-19 untuk lansia dan bukan lansia (semua warga) di Puskesmas Kecamatan Kramat Jati dengan hanya membawa e-KTP asli dan memakai masker dengan protokol 3M, adalah tidak benar. Faktanya, Puskesmas Kecamatan Kramat Jati tidak mengeluarkan pengumuman tersebut dan pesan tersebut dibuat oleh oknum yang tidak bertanggung jawab.	Fabricated Content	3
Beredar pesan berantai melalui aplikasi WhatsApp yang menginformasikan agar sertifikat vaksin Covid-19 yang diterima via WhatsApp harus disimpan untuk menghindari kesalahan pemberian jenis vaksin saat vaksin yang kedua. Hal ini dikarenakan tim medis tidak akan mengingat jenis/ tipe vaksin yang sudah diberikan penerima vaksin, sedangkan pemberian jenis vaksin kedua harus sama dengan vaksin yang pertama.	Berdasarkan hasil koordinasi Tim Jalahoaks dengan Dinas Kesehatan Provinsi DKI Jakarta (04/03/2021), diperoleh klarifikasi bahwa pesan tersebut keliru. "Di sistem pencatatan pelaporan online Pcare nya sudah ada info tanggal, nomor batch sampai merek vaksin setiap penerima. Jadi pesan yang beredar hoaks ya," kata Staf Dinas Kesehatan Provinsi DKI Jakarta saat dihubungi Tim Jalahoaks. Dilansir dari website resmi Kementerian Kesehatan RI sehatnegeriku.kemkes.go.id (21/01/2021), aplikasi Pcare vaksin Covid-19 merupakan bagian dari sistem informasi satu data vaksinasi Covid-19. Pcare mendukung proses registrasi sasaran penerima vaksin, screening status kesehatan, serta mencatat dan melaporkan hasil pelayanan vaksinasi Covid-19.	Informasi bahwa sertifikat vaksin Covid-19 yang diterima via WhatsApp harus disimpan untuk menghindari kesalahan pemberian jenis vaksin saat vaksin yang kedua karena tim medis tidak akan mengingat jenis/ tipe vaksin yang sudah diberikan, adalah tidak benar. Faktanya, informasi tanggal, nomor batch , sampai merek vaksin setiap penerima vaksin Covid-19 telah tercatat pada aplikasi Pcare, yakni sistem informasi satu data vaksinasi Covid-19.	Fabricated Content	3

Sample data on Table 2 consist of 7 label category of the hoax that divided into 3 class. The class division of this hoax category based on the manipulation and purpose of the hoax news. Satire or Parody, False Connection, Misleading Content, categorized into class 1. False Context, Imposter Content categorized into class 2 and Manipulated Content, Fabricated Content categorized into class 3. This data can access on bit.ly/32fsXP5.

2.2 Text Processing and Stemming Nazief & Adriani

Text preprocessing changed unstructured data into structured [17]. Steps of text preprocessing are case folding, tokenization, filtering, and stemming. Stemming applied the Nazief & Adriani algorithm stemming technique [18] apply the following rules:

1. The word check and match in the root word dictionary; if found, the process will stop, but then the following process will continue.
2. Removing inflectional suffix {"-kah," "-lah," "-tah," "-pun"} and suffix {"-ku," "-mu," or "-nya"} with rule [[[DP+]DP+]DP+] root-word [+DS].
3. Removing derivational suffix y {"-i," "-kan," and "-an"} with rule [[[DP+]DP+]DP+] root-word.
4. Removing derivational prefix {"be-," "di-," "ke-," "me-," "pe-," "se-," and "te-"}.
5. The process of re-checking the word by removing the prefix by changing it to the rules and re-checking the root word dictionary if it is still not found, then the following process will be carried out.
6. If all steps have been completed and no results are found, the word would be considered the root word, and the initial word value will be returned.

2.3 KNN in Jaccard Space

KNN is a model that classifies objects based voting on a given collection [19]. This classifier algorithm also works by initially determining the distance in Equation (1), sorting by nearest K distance and using the majority voting of the K parameter [7].

$$D(X,Y) = \sqrt{d_1(x_1,y_1) + d_1(x_1,y_1) + \dots\dots\dots + d_n(x_n,y_n)} \quad (1)$$

Distance in Equation (1) defined based on two point data. Jaccard is the most commonly used distance in data to know the similarity between two sets. Let A and B be two sets. Jaccard index is the sliced population compared to all items in both sets [8].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Equation (2) shows Jaccard Distance between untersect two sets A and B also belong to member B. Example q-gram Jaccard distance calculation as below:

Sentence A = "Covid 19 adalah virus yang mematikan"
Sentence B = "Covid 19 virus yang sangat berbahaya"

Set of q-gram A: {"a", "c", "d", "e", "g", "h", "i",…….., "ati", "tik", "ika", "kan"}
Set of q-gram B: {"a", "b", "c", "e", "g",……, "bah", "aha", "hay", "aya"}
with assumption: regardless their order, without duplicating character combinations in order to avoid typos.

A∩B= {"a", "c", "e", "g", …. , "iru", "rus", "yan", "ang"} ➔ 41 character combination
A∪B={"a", "b", "c", "e", "g", ……,"aha", "hay", "aya"} ➔ 113 character combination

J(A,B) = (41)/(113) = 0.36
The similarity of A and B is 1 - 0.36 = 0.64

Furthermore, before looking for the distance between the data and the neighbors, determine value of K. The KNN model in this research will select the K parameter with odd number are 3, 5 and 7 [6],[7].

2.4 Evaluation Model

Similar to other researchers about hoax detection, this reaseach focus on evaluation of Accuracy. Accuracy is used to evaluate the number of predictive labels that correspond to the actual label [20]. Accuracy obtained from confusion matrix in Figure 1.

		Predicted Class P	Predicted Class N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 1 Confusion Matrix

True Positive (TP) is positive data that predicted to be correct, True Negative (TN) is negative data that predicted to be correct, False Positive (FP) or Type I Error is negative data but predicted to be positive data., and False Negative (FN) or Type II Error is positive data but predicted as negative data. So, equation of acuracy can shown in Equation (3).

$$Accuracy = \frac{TN + TP}{(TP + FP + TN + FN)} \quad (3)$$

Accuracy used as a reference for algorithm performance, if dataset has a close number of FN and FP.

3. RESULTS AND DISCUSSION

3.1 Exploration

Preprocessing text eliminate words to reduce noise from dataset. The results of the word weighting TF-IDF can be seen in wordcloud in Figure 2 and scatter word shown in Figure 3.

Figure 2 Wordcloud related Covid-19 Hoax

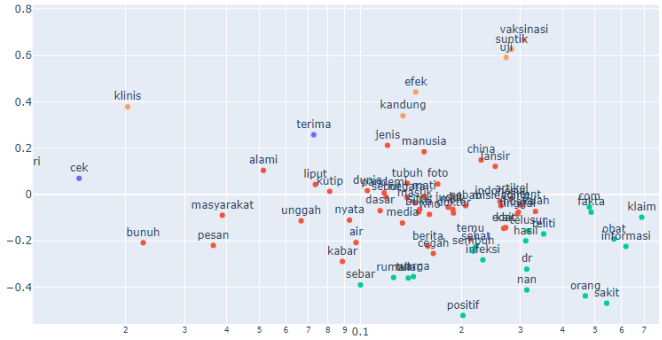

Figure 3 Scatter Word realated Covid-19 Hoax

Wordcloud on Figure 2 shows words often used for hoax-related Covid-19 are “Virus”, “Covid”, “Vaksin”, “Fakta”, “Video”, “Klaim”, “Corona”, and others. These words have been selected by TF-IDF. Scatter Word in Figure 3 describe the layout of words that are often used in several sentences. As an example of an orange cluster, there are several words, namely “vaksinasi”, “suntik” dan “uji”. It can be seen that hoaxes in this cluster are more dominated by topics related to vaccines from Covid-19 or effects after vaccine injection.

3.2 Classification

The scenario of this classification was done using two approaches, namely without stemming and stemming Nazief & Adriani. This classification label consists of three labels, namely, Class_1, Class_2, and Class_3.

3.2.1 Without Stemming Nazief & Adriani

Classification without stemming Nazief & Adriani performs classification based on TF-IDF and words (Jaccard similarity) without normalizing stemmed from Nazief & Adriani. The performances of classification without stemming Nazief & Adriani are shown in Table 3.

Table 3 Classification without Stemming Nazief & Adriani

Model	Distance	Accuracy	Precision	Recall	F1-Score
KNN with K=3	Euclidean	52,68%	43,79%	39,97%	39,13%
	Manhattan	64,29%	59,76%	56,23%	57,30%
	Minkowski	52,68%	43,79%	39,97%	39,13%
	Jaccard	69,64%	70,73%	58,86%	61,65%
KNN with K=5	Euclidean	58,04%	62,31%	45,33%	46,46%
	Manhattan	60,71%	59,35%	48,75%	48,88%
	Minkowski	58,04%	62,31%	45,33%	46,46%
	Jaccard	**69,64%**	**73,61%**	**60,28%**	**62,80%**
KNN with K=7	Euclidean	58,93%	53,97%	44,92%	44,06%
	Manhattan	63,39%	60,31%	52,47%	53,75%
	Minkowski	58,93%	53,97%	44,92%	44,06%
	Jaccard	67,86%	72,12%	56,16%	59,29%
Naïve Bayes		66,07%	70,45%	49,86%	49,94%

Table 3 shows the best accuracy is KNN with Jaccard Space is 69,64% with precision 73,61%. Best recall and f1-score is KNN with Jaccard Space with 60,28% and 62,80%. Classification hoax related Covid-19 without Stemming Nazief & Adriani get the best performance on KNN model in Jaccard space. The difference of accuracy Naïve Bayes with KNN model with Jaccard Space is not far below 5%.

3.2.2 With Stemming Nazief & Adriani

Model and distance are the same with scenario before. This scenario compare the effect of stemming Nazief & Adraini. Performances classification with stemming Nazief & Adriani are shown in Table 4.

Table 4 Classification Result with Stemming Nazief & Adriani

Model	Distance	Accuracy	Precision	Recall	F1-Score
KNN=3	Euclidean	58,93%	55,16%	45,40%	45,66%
	Manhattan	58,04%	53,26%	53,45%	53,33%
	Minkowski	58,93%	55,16%	45,40%	45,66%
	Jaccard	69,64%	71,93%	60,37%	63,31%
KNN=5	Euclidean	54,46%	51,18%	40,50%	39,80%
	Manhattan	60,71%	57,26%	51,96%	52,32%
	Minkowski	54,46%	51,18%	40,50%	39,80%
	Jaccard	**75,89%**	**79,55%**	**67,50%**	**71,13%**
KNN=7	Euclidean	50,89%	39,85%	36,71%	34,81%
	Manhattan	62,50%	55,53%	52,01%	51,92%
	Minkowski	50,89%	39,85%	36,71%	34,81%
	Jaccard	74,11%	76,66%	64,32%	67,70%
Naïve Bayes		65,18%	73,40%	48,18%	48,33%

Based on Table 4 above, the best performance classification obtained by KNN in Jaccard Distance with K=5 with 75.89% accuracy with 79,55% precision, 67,50% recall and 71,13% f1-score. Improvement Stemming Nazief & Adriani from first scenario (without stemming) to second scenario (with stemming) obtained 6,25% for KNN in Jaccard Space in K=5.

3.3 Discussion

The accuracy of KNN with Jaccard Space has been improved on second scenario. This is due to break the words into a combination of characters and good to solve typos in the sentence. Best accuracy for first scenario is 69.64%, while second scenario is 75.89% or difference 6.25% from first scenario of KNN in Jaccard space with K = 5. Confusion matrix on KNN in Jaccard Space with K = 5 for second scenario is shown in Figure 4.

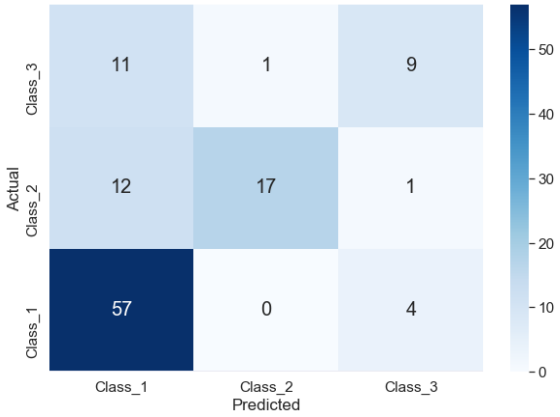

Figure 4 Confusion Matrix

Figure 4. explain that dominance of Class_1, which was successfully predicted to be 57 correct with 12 wrong on Class_2 and 11 on Class_3. While, Class_2 was successfully predicted 17 and incorrectly predicted 1 in Class_3. Finally, Class_3 was correctly predicted 9 and incorrectly predicted 1 in Class_2 and 4 in Class_1. Testing carried out on the data on the

train data for example with sentence "vaksin covid-19 ditanami barcode yang akan masuk dalam tubuh manusia". The results of the testing is shown in Table 5.

Table 5. Result Prediction.

No.	Model	Class Category
1	KNN in Jaccard Space (K=5)	2
2	Naïve Bayes	1

The hoax categories in Table 5, which is predicted in the KNN in the Jaccard Space model with K = 5, is 2, Naïve Bayes is 1. In this prediction, the KNN in the Jaccard Space model with K = 5 correctly predicts the class category. This shows that the hoax against "vaksin covid-19 ditanami barcode yang akan masuk dalam tubuh manusia" is included in the hoax category of Misleading Content or False Context or Imposter Content.

The improvement of this research compared to previous research is distance and stemming used, while for other parameters it is almost same as the previous research.

Table 6 Comparison of Perfomance.

Parameter	**Base Scenario**	**Improved Scenario (This Research)**
Distance	KNN with Euclid [12] obtained average accuracy of 55,66%	KNN with Jaccard Space obtained average accuracy of 71,13%
With Nazief & Adriani Stemming	KNN with Euclidean distance TF-IDF [6] obtained average accuracy of 54,76%	K-NN in Jaccard Space obtained average accuracy of 73,21%
Without Nazief & Adriani Stemming	KNN with Euclidean distance TF-IDF [7] obtained average accuracy of 56,55%	K-NN in Jaccard Space obtained average accuracy of 69,05%

Base scenario for distance parameter for Euclid and Jaccard have diffrence 15,48% in accuracy. Furthermore, parameters without Nazief & Andriani Stemming increase 12.5% in accuracy, while parameters with Naizef & Andrian Stemming increase 18.45% in accuracy.

4. CONCLUSIONS

This research was done to classify hoax news on Covid-19 using KNN in Jaccard Space. This model improved previous models such as KNN with Euclidean, Manhattan, and Minoski distance. KNN in Jaccard Space model with K = 5 was better in accuraucy of 75.89%, than 65.18%. Handling hoax news is about blocking, but more importantly, in the long term, preparing people to understand hoax news. Several activities that can be spared hoax news pay attention to news sources (the site is trusted) and fact-checking from several media on the internet. Further research can combine several sources from Instagram, Twitter, or Facebook to maximize current hoaxes on social media.

REFERENCES

[1] C. Juditha, "Hoax Communication Interactivity in Social Media and Anticipation (Interaksi Komunikasi Hoax di Media Sosial serta Antisipasinya)," *J. Pekommas*, vol. 3, no. 1, p. 31, 2018, doi: 10.30818/jpkm.2018.2030104.

[2] C. Juditha, "People Behavior Related To The Spread Of Covid-19's Hoax," *J. Pekommas*, vol. 5, no. 2, p. 105, 2020, doi: 10.30818/jpkm.2020.2050201.

[3] J. S. Hoaks, "Jabar Saber Hoaks," *Instagram*. 2020, [Online]. Available: https://www.instagram.com/jabarsaberhoaks/ accesed on 06 June 2020.

[4] B. Zaman, A. Justitia, K. N. Sani, and E. Purwanti, "An Indonesian Hoax News Detection System Using Reader Feedback and Naïve Bayes Algorithm," *Cybern. Inf. Technol.*, vol. 20, no. 1, pp. 82–94, 2020, doi: 10.2478/cait-2020-0006.

[5] A. Fauzi, E. B. Setiawan, and Z. K. A. Baizal, "Hoax News Detection on Twitter using Term Frequency Inverse Document Frequency and Support Vector Machine Method," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012025.

[6] R. Sagita, U. Enri, and A. Primajaya, "Klasifikasi Berita Clickbait Menggunakan K-Nearest Neighbor (KNN)," *JOINS (Journal Inf. Syst.*, vol. 5, no. 2, pp. 230–239, 2020, doi: 10.33633/joins.v5i2.3705.

[7] A. Kesarwani, S. S. Chauhan, and A. R. Nair, "Fake News Detection on Social Media using K-Nearest Neighbor Classifier," *Proc. 2020 Int. Conf. Adv. Comput. Commun. Eng. ICACCE 2020*, pp. 0–3, 2020, doi: 10.1109/ICACCE49060.2020.9154997.

[8] S. Sunardi, A. Yudhana, and I. A. Mukaromah, "Implementasi Deteksi Plagiarisme Menggunakan Metode N-Gram Dan Jaccard Similarity Terhadap Algoritma Winnowing," *Transmisi*, vol. 20, no. 3, p. 105, 2018, doi: 10.14710/transmisi.20.3.105-110.

[9] E. Y. Sari, A. D. Wierfi, and A. Setyanto, "Sentiment Analysis of Customer Satisfaction on Transportation Network Company Using Naive Bayes Classifier," *2019 Int. Conf. Comput. Eng. Network, Intell. Multimedia, CENIM 2019 - Proceeding*, vol. 2019-Novem, 2019, doi: 10.1109/CENIM48368.2019.8973262.

[10] A. P. Ardhana, D. E. Cahyani, and Winarno, "Classification of Javanese Language Level on Articles Using Multinomial Naive Bayes and N-Gram Methods," *J. Phys. Conf. Ser.*, vol. 1306, no. 1, pp. 0–9, 2019, doi: 10.1088/1742-6596/1306/1/012049.

[11] H. Mustofa and A. A. Mahfudh, "Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes," *Walisongo J. Inf. Technol.*, vol. 1, no. 1, p. 1, 2019, doi: 10.21580/wjit.2019.1.1.3915.

[12] K. Umam, "Group chat analysis of hoax detection during the covid-19 pandemic using the k nearest neighbors algorithm and massive text processing," *J. Phys. Conf. Ser.*, vol. 1918, no. 4, p. 042149, 2021, doi: 10.1088/1742-6596/1918/4/042149.

[13] M. A. Rahmat, Indrabayu, and I. S. Areni, "Hoax web detection for news in bahasa using support vector machine," *2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019*, pp. 332–336, 2019, doi: 10.1109/ICOIACT46704.2019.8938425.

[14] B. Irena and Erwin Budi Setiawan, "Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 4, pp. 711–716, 2020, doi: 10.29207/resti.v4i4.2125.

[15] S. Temma, M. Sugii, and H. Matsuno, "The Document Similarity Index based on the Jaccard Distance for Mail Filtering," *34th Int. Tech. Conf. Circuits/Systems, Comput. Commun. ITC-CSCC 2019*, pp. 3–6, 2019, doi: 10.1109/ITC-CSCC.2019.8793419.

[16] P. Sangwan and R. Behl, "Truth Detection in Social Media Posts using Jaccard Algorithm with SRTD and Word Net Concept," *Proc. Int. Conf. Res. Manag. Technovation 2020*, vol. 24, pp. 103–107, 2020, doi: 10.15439/2020km24.

[17] T. Winarti, J. Kerami, and S. Arief, "Determining Term on Text Document Clustering using Algorithm of Enhanced Confix Stripping Stemming," *Int. J. Comput. Appl.*, vol. 157, no. 9, pp. 8–13, 2017, doi: 10.5120/ijca2017912761.

[18] J. Asian, H. E. Williams, and S. M. M. Tahaghoghi, "Stemming Indonesian," *Conf. Res. Pract. Inf. Technol. Ser.*, vol. 38, no. 4, pp. 307–314, 2005, doi: 10.1145/1316457.1316459.

[19] T. Granskogen and J. A. Gulla, "Fake news detection: Network data from social media used to predict fakes," *CEUR Workshop Proc.*, vol. 2041, no. 1, pp. 59–66, 2017.

[20] A. F. Iskandar, E. Utami, and A. B. Prasetio, "Word Analysis of Indonesian Keirsey Temperament," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, Vol. 14, No. 4, pp. 365–376, 2020. doi: d10.22146/ijccs.58595.